Leaner and meaner genomes in Escherichia coli

Ussery, David

Published in:
Genome Biology

Link to article, DOI:
10.1186/gb-2006-7-10-237

Publication date:
2006

Document Version
Publisher's PDF, also known as Version of record

Minireview

Leaner and meaner genomes in *Escherichia coli*

David W Ussery

Address: Center for Biological Sequence Analysis, BioCentrum-DTU, Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark.
Email: dave@cbs.dtu.dk

Published: 24 October 2006

Genome ***Biology*** 2006, **7:**237 (doi:10.1186/gb-2006-7-10-237)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2006/7/10/237

Abstract

A 'better' *Escherichia coli* K-12 genome has recently been engineered in which about 15% of the genome has been removed by planned deletions. Comparison with related bacterial genomes that have undergone a natural reduction in size suggests that there is plenty of scope for yet more deletions.

Why should one want to design a better bacterium? One answer is that this is one way of really testing our understanding of how a living cell works - by making predictions, manipulating the object, and seeing what we get. This is the province of synthetic biology, whose ultimate goal is to understand life by constructing it from scratch; it is hoped that along the way will emerge an understanding of the properties of living cells and organisms that is difficult to arrive at by conventional investigation of the organisms themselves [1,2]. Much progress has been made recently towards designing and synthesizing novel biological organisms from a set of standardized parts [3], such as protein-coding genes, regulators, and terminators, as listed on the BioBrick website [4].

In contrast, in work recently published in *Science*, Posfai *et al.* [5] have taken a 'deconstructionist' approach to redesigning life. Specific regions of the *Escherichia coli* K-12 genome were targeted for deletion with the intention of improving the production potential of this model organism. As an unanticipated side effect, they have come up with a bacterium that is even better than the parental strain for some purposes, in that it is more efficiently electroporated and accurately propagates unstable recombinant genes and plasmids. It is interesting to compare these engineered reduced genomes [5] with the genomes of other bacteria within the Enterobacteriaceae, some of which are endosymbionts whose genomes have become dramatically reduced during evolution.

Smaller is indeed often better, as people who travel frequently or who worry about buying fuel for their cars know. Posfai *et al.* [5] chose which genes and genomic regions to remove on the basis of several criteria, including "troublesome sequences" such as insertion sequence (IS) sites and transposable elements that appear to code only for their own replication, and repeat regions that can cause homologous recombination. They also removed some regions that are not present in all *E. coli* genomes, and so are unlikely to be essential for basic properties such as growth. There are many large regions throughout the *E. coli* K-12 genome that are not conserved among other *E. coli* genomes, but given the variation in genome size between different strains, with differences of more than 1 million base pairs (20% of the genome) being common, this is perhaps not surprising.

To make the deletions, synthetic oligomers containing regions homologous to target sites flanking the desired region were used. Regions were deleted by recombination mediated by the phage lambda Red system, and done in a way that gave 'scarless' deletions where no marker sequence was left. The strains with deletions were then tested for growth in minimal media. Finally, as one last step to check for quality, the reduced strains were hybridized to tiling microarrays of the *E. coli* K-12 genome. The first reduced strain yielded surprising results. In the words of the authors: "Alarmingly, we found five unexpected copies of IS that had transposed to new locations since the project began in 2002." Thus new strains were made, which were shown to be free of IS elements. The engineered strains had similar growth rates to their parent strain, and the electroporation efficiency of engineered strain MDS42 was 100 times greater than for the original *E. coli* MG1655 K-12. Furthermore, plasmid genes

Table 1

List of currently sequenced genomes from the family Enterobacteriaceae of the γ-Proteobacteria

Number of proteins	Genome size (bp)	Organism	%AT	Number of tRNA genes	Number of rRNA genes	Accession number
5,379	**5,231,428**	***Escherichia coli* CFT073**	**49.5**	**89**	**7**	**AE014075**
5,361	**5,498,450**	***Escherichia coli* O157 RIMD**	**49.5**	**105**	**7**	**BA000007**
5,349	**5,528,445**	***Escherichia coli* O157 EDL**	**49.5**	**98**	**7**	**AE005174**
5,066	**5,065,741**	***Escherichia coli* UTI89**	**49.4**	**88**	**7**	**CP000243**
4,905	5,688,987	*Photorhabdus luminescens*	57.2	85	7	AP009048
4,685	**4,938,920**	***Escherichia coli* strain 536**	**49.5**	**81**	**7**	**CP000247**
4,600	4,809,037	*Salmonella enterica* CT18	47.9	79	7	AL513382
4,492	5,064,019	*Erwinia carotovora*	49.0	76	7	BX950851
4,452	4,857,432	*Salmonella typhimurium* LT2	47.8	85	7	AE006468
4,445	4,755,700	*Salmonella enterica* SCB67	47.8	85	7	AE017220
4,436	4,607,203	*Shigella flexneri* 2a301	49.1	97	7	AE005674
4,337	**4,646,332**	***Escherichia coli* K-12 W3110**	**49.2**	**86**	**7**	**U00096**
4,331	**4,639,675**	***Escherichia coli* K-12 MG1655**	**49.2**	**86**	**7**	**AP009048**
4,323	4,791,961	*Salmonella enterica* Ty2	47.2	78	7	AE014613
4,277	4,369,232	*Shigella dysenteriae* Sd197	48.8	85	7	CP000034
4,224	4,825,265	*Shigella sonnei* Ss046	49.0	97	7	CP000038
4,167	4,702,289	*Yersinia pestis* Antiqua	52.3	68	7	CP000308
4,142	4,519,823	*Shigella boydii* Sb227	48.8	91	7	CP000036
4,116	4,574,284	*Shigella flexneri* 5str8401	49.1	91	7	CP000266
4,093	4,585,229	*Salmonella enterica* ATCC9150	47.8	82	7	CP000026
4,090	4,600,755	*Yersinia pestis* KIM	52.4	73	7	AE009952
4,073	4,599,354	*Shigella flexneri* 2457T	49.1	98	7	AE014073
4,011	**4,263,492**	***Escherichia coli* MDS12**	**49.2**	**86**	**7**	**[5]**
4,008	4,653,728	*Yersinia pestis* CO-92	52.4	70	6	AL590842
3,981	4,534,590	*Yersinia pestis* Nepal516	52.4	72	7	CP000305
3,974	4,744,671	*Yersinia pseudotuber* IP32953	52.4	85	7	BX936398
3,895	4,595,065	*Yersinia pestis* Mediaevails	52.3	72	7	AE017042
3,731	**3,977,067**	***Escherichia coli* MDS41**	**49.2**	**86**	**7**	**[5]**
3,730	**3,976,359**	***Escherichia coli* MDS42**	**49.2**	**86**	**7**	**[5]**
3,691	**3,931,408**	***Escherichia coli* MDS43**	**49.2**	**86**	**7**	**[5]**
2,432	4,171,146	*Sodalis glossinidius*	45.3	69	7	AP008232
611	697,724	*Wiggelsworthia glossinidia*	77.5	34	2	BA000021
610	791,654	*Blochmannia pennsylvanicus*	70.4	39	1	CP000016
595	686,194	*Baumannia cicadellinicola*	61.2	39	2	CP000238
589	705,557	*Blochmannia floridanus*	72.6	37	1	BX248583
564	640,681	*Buchnera aphidicola* APS	73.7	32	1	BA000003
545	641,454	*Buchnera aphidicola* Sg	74.7	32	1	AE013218
504	615,980	*Buchnera aphidicola* BBp	74.7	32	1	AE016826

The genomes are sorted by the number of genes, from the largest to smallest, and the *E. coli* genomes are in bold. The multi-deletion series (MDS) strain data are from Posfai *et al.* [5], and data for the other genomes are from EMBL files (EMBL accession numbers in the last column). An up-to-date table of the enteric genomes sequenced and available so far can be obtained from the GenomeAtlas database [13] by typing in 'BProt GE' in the keyword search.

that were unstable in MG1655 were found to be completely stable in the engineered strains. IS mutagenesis is a natural defense against deleterious genes, and is normally helpful to bacteria in the wild, but is detrimental when one wishes to grow these genes in laboratory strains of *E. coli*.

Natural genome reduction

If synthetic biology can be used to design a reduced *E. coli* genome with some desirable new functions, what about 'non-synthetic' biology - that is, evolution? Is there anything that we can learn from evolutionary biology about how to make a

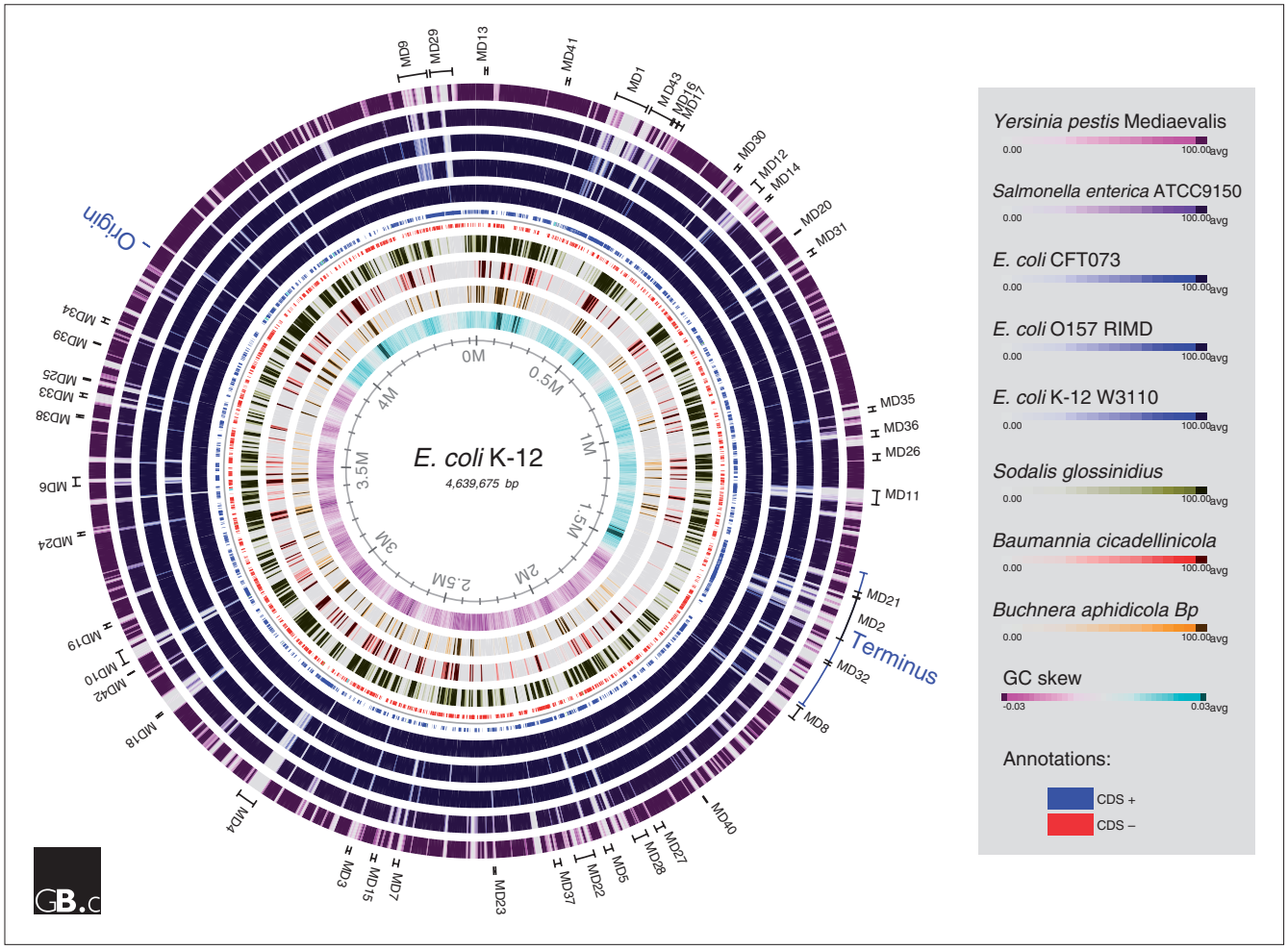

Figure 1
A BLAST atlas diagram of eight enteric bacterial genomes, compared to the reference *E. coli* K-12 isolate MG1655. The outer five circles are other similar genomes from *E. coli*, *S. enterica* and *Y. pestis* (outer circle), while the inner three circles reflect the reduced genomes (from the innermost circle outwards) of *B. aphidicola*, *B. cicadellinicola* and *S. glossinidius*. Each colored circle represents the BLAST score of the best hit of the given bacterial proteome versus the gene at a given location in the reference *E. coli* MG1655 genome. Note that the scale is -log E-value, which means that the strongly colored regions have an E-value of less than 10^{-100}, which corresponds to a very good match. The locations of all the deletions engineered by Posfai *et al.* [5] are indicated outside the circles. Coding sequences of the reference genome (*E. coli* k-12 strain MG1655) are indicated as blue and red blocks, corresponding to genes orientated clockwise or counterclockwise. The gaps in the *E.coli* W3110 genome are due to rRNA operons and other non-coding RNAs, which do not show up on the protein BLAST results.

reduced *E. coli* genome? Soon after Posfai *et al.* [5] published their paper, a study by Wu *et al.* [6] appeared on the reduced genomes of two very different bacteria living inside an insect called the glassy-winged sharpshooter (*Homalodisca coagulata*) [6]. One of these bacteria (*Baumannia cicadellinicola*) belongs phylogenetically to the same group as *E. coli*, which inspired me to make a comparison of all of the engineered *E. coli* genomes of Posfai *et al.* [5] with other enteric bacterial genomes sequenced so far (Table 1).

The *B. cicadellinicola* genome is towards the bottom of the table, but there are four known genomes in this family that encode an even smaller number of proteins. The genome at the bottom of the list (*Buchnera aphidicola* strain BBp) codes for only 504 proteins, or less than 10% of the number of proteins encoded by the larger *E. coli* genomes (5,379 proteins in *E. coli* CFT073). The smallest 'normal', free-living enterobacter (apart from the newly engineered *E. coli* genomes) is a *Yersinia pestis* strain (Mediaevalis), with 3,895 genes, or only 72% of the number of genes found in the largest enterobacterial genome. Furthermore, only slightly more than half of the *Y. pestis* Mediaevalis genes have homologs in the CFT073 genome (52% - that is, 2,938 *Y. pestis* genes/5,379 *E. coli* CFT073 genes). Thus, just on the basis of gene diversity within the enteric bacteria, it seems that perhaps half or more of the genes in the larger enterobacterial genomes might be expendable - at least under laboratory growth conditions. Indeed, only 620 *E. coli* K-12 genes have been found experimentally to be essential for growth in rich media, while 3,126 genes were found to be

dispensable for growth under this condition [7]. This indicates that there could well be more room for engineered deletions in the *E. coli* genomes.

Figure 1 compares the proteins encoded by some of the genomes in Table 1. *E. coli* K-12 isolate MG1655 was the genome used as a starting point by Posfai *et al.* [5] to make the reduced genomes, and it is used as the reference here. The locations of all the deletions engineered by Posfai *et al.* [5] are indicated outside the circles. The location of each gene from the *E. coli* MG1655 genome is used to visualize the best hit based on a BLAST search of all the proteins encoded by the different genomes indicated in the legend. A strong hit is represented by a solid bar so the extensive solid regions in the circles represent regions of homology with the other genomes. Each circle represents a different genome, with the outer five circles representing nonreduced genomes, and the inner three circles reduced genomes. Thus, the outermost circle is for the *Y. pestis* genome, followed by a *Salmonella* genome, and then three *E. coli* genomes. Note that many of the planned deletions made by Posfai *et al.* [5] correspond to gaps in the otherwise mostly solid dark circles; these gaps represent large chromosomal regions lacking homology in terms of protein sequences in the other genomes.

The reduced genomes are shown in the inner circles. *B. aphidicola* strain BBp is the smallest genome, and is depicted as the orange inner circle, which has few hits, as expected, as this genome encodes so few proteins. The *B. cicadellinicola* genome is the next circle (red), and the third is *Sodalis glossinidius*, which is a genome that is undergoing reduction, but still contains about 2,500 genes, as well as about 1,000 pseudogenes [8]. This circle contains more hits, although it is still a bit sparse compared to the inner three circles, which have large regions where nearly all of the proteins are conserved.

These reduced genomes contain only about 10% of the genes in the larger *E. coli* genomes from which they originated long ago. This raises many questions. What about the remaining 90%? Does *E. coli* really not need most of these genes? Some are certainly redundant - a necessary condition for robust systems [9] - and the definition of 'essential genes' might include some genes that do not give a lethal phenotype when deleted [10].

Is it possible to model which genes would remain, and which 90% or so could be removed, under the right conditions? A model of *E. coli* metabolism was recently used to generate reduced genomes *in silico* [11], and to compare these genomes with the endosymbiotic genomes shown at the bottom of Table 1. The idea was to use a known metabolic environment and then to model random gene loss, and evaluate relative viability. If the gene loss had no apparent effect, then another gene would be removed, and this process was repeated until a minimal genome was obtained. Two different endosymbiotic bacterial environments were modeled - those of *Buchnera* and *Wiggelsworthia* - and the model predicted the gene content of the two genomes to about 80% accuracy [11].

There are, of course, several different ways to arrive at the same reduced genome, but by looking at which genes are necessary to perform core metabolic activities (for a given endosymbiotic environment, it should be stressed), it is possible in general to predict the genes that are likely to remain in a reduced genome. This information can then be used in future experiments to design better genomes, tailor-made for specific applications. Posfai *et al.* [5] were not intending to manufacture a 'minimal genome' such as the highly reduced ones discussed here, but rather they simply wanted to engineer an *E. coli* genome that would be a better 'workhorse' - that is, it would be easier to get DNA into the cells, and the DNA and its gene products would be stable once it was there. There are others, however, who do have the aim of using synthetic biology to design and manufacture a minimal genome [12]. Perhaps the time is near when mircobiology will join the engineering sciences.

Acknowledgements

I would like to thank Michael Sismour for useful comments about synthetic biology and the Danish Center for Scientific Computing for funding.

References

1. Benner SA, Sismour AM: **Synthetic biology.** *Nat Rev Genet* 2005, **6:**533-543.
2. de Lorenzo V, Serrano L, Valencia, A: **Synthetic biology: challenges ahead.** *Bioinformatics* 2006, **22:**127-128.
3. Fu P: **A perspective of synthetic biology: assembling building blocks for novel functions.** *Biotechnol J* 2006, **1:**690-699.
4. **The BioBricks Foundation** [http://openwetware.org/wiki/The_BioBricks_Foundation]
5. Posfai G, Plunkett G 3rd, Feher T, Frisch D, Keil GM, Umenhoffer K, Kolisnychenko V, Stahl B, Sharma SS, de Arruda M, *et al.*: **Emergent properties of reduced-genome *Escherichia coli*.** *Science* 2006, **312:**1044-1046.
6. Wu D, Daugherty SC, Aken SE, Pai GH, Watkins KL, Khouri H, Tallon LJ, Zaborsky JM, Dunbar HE, Tran PL, *et al.*: **Metabolic complementarity and genomics of the dual bacterial symbiosis of sharpshooters.** *PLoS Biol* 2006, **6:**e188.
7. Gerdes SY, Scholle MD, Campbell JW, Balazsi G, Ravasz E, Daugherty MD, Somera AL, Kyrpides NC, Anderson I, Gelfand MS, *et al.*: **Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655.** *J Bacteriol* 2003, **185:**5673-5684.
8. Toh H, Weiss BL, Perkin SAH, Yamashita A, Oshima K, Hattori M, Aksoy S: **Massive genome erosion and functional adaptations provide insights into the symbiotic lifestyle of *Sodalis glossinidius* in the tsetse host.** *Genome Res* 2006, **16:**149-156.
9. Kafri R, Levy M, Pilpel Y: **The regulatory utilization of genetic redundancy through responsive backup circuits.** *Proc Natl Acad Sci USA* 2006, **103:**11653-11658.
10. Fang G, Rocha E, Danchin A: **How essential are nonessential genes?** *Mol Biol Evol* 2005, **22:**2147-2156.
11. Pal C, Papp B, Lercher MJ, Csermely P, Oliver SG, Hurst LD: **Chance and necessity in the evolution of minimal metabolic networks.** *Nature* 2006, **440:**667-670.
12. Smith HO, Hutchison CA 3rd, Pfannkoch C, Venter JC: **Generating a synthetic genome by whole genome assembly: phiX174 bacteriophage from synthetic oligonucleotides.** *Proc Natl Acad Sci USA* 2003, **100:**15440-15445.
13. **CBS Genome Atlas Database** [http://www.cbs.dtu.dk/services/GenomeAtlas]